comparison web kids

by Nur Hayatin

Submission date: 01-Jan-2020 10:26PM (UTC+0700)
Submission ID: 1238926576
File name: Sentra_2019_-_klasifikasi_web_anak.pdf (659.6K)
Word count: 1693
Character count: 9900

Seminar Nasional Teknologi dan Rekayasa (SENTRA) 2019
ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

COMPARISON OF EXTRACTION KEYWORD RESULT FROM WEB CATEGORY OF KID AND GENERAL

Nur Hayatin[1], Gita Indah Marthasari[2], Maulidya Yuniarti[3]
[123]Universitas Muhammadiyah Malang, Malang

Kontak Person:
Nur Hayatin
Jalan Raya Tlogomas No. 246, Tlogomas, Lowokwaru, Kota Malang, Jawa Timur 65144
E-mail: noorhayatin@umm.ac.id

Abstrak

Kebutuhan akan ekstraksi konten web yang spesifik untuk kategori anak cukup tinggi. Diantaranya digunakan untuk konten filtering, klasifikasi web anak, dan pencarian web anak. Ekstraksi kata kunci adalah salah satu metode yang dapat digunakan untuk melakukan pemilihan kata yang bertujuan untuk mendapatkan informasi terkait semantik. Permasalahannya belum ada penelitian yang fokus pada ekstraksi kata kunci dengan topik spesifik untuk web anak. Penelitian ini telah melakukan ekstraksi kata kunci pada web anak berdasarkan word frequency. Kami melakukan crawling data dari directory web Dmoz sebanyak lebih dari 1000 link url untuk kategori general, kid, teens, dan mature teens. Dari hasil 100 kata yang memiliki nilai word frequency teratas untuk masing-masing kategori tersebut, kami menemukan bahwa ada perbedaan signifikan untuk kata kunci yang muncul pada web kategori anak dan general.

Kata kunci: *keyword extraction, web category, term frequency, web kids.*

1. Pendahuluan

Dengan pesatnya perkembangan Internet, informasi yang terkandung dalam halaman Web meningkat secara eksplosif. Saat menelusuri halaman Web, orang sering dengan mudah mengabaikan informasi berharga hanya dengan tergantung pada judul halaman dan hasil pencarian. Menghadapi web besar-besaran, bagaimana memahami konten utama dari halaman-halaman besar yang dibantu oleh teknologi informasi telah menjadi pusat penelitian. Cara efektif untuk mengatasi masalah ini adalah mengekstrak kata kunci dari halaman Web **[1]**. Ekstraksi kata kunci (Keyword extraction) adalah topik penelitian penting dalam information retrieval. Yang dimaksud kata kunci dalam sebuah dokumen biasanya merujuk pada beberapa kata atau frasa yang paling terkait dengan konten dokumen tersebut. Ekstraksi kata kunci mengacu pada pemilihan kata-kata fitur dari teks. Kata kunci dapat memberikan informasi semantik untuk banyak aplikasi text mining, misalnya : klasifikasi dokumen, klasterisasi, retrieval, analisis dan pencarian berbasis topik **[2]**.

Banyak penelitian telah dilakukan untuk penggalian kata kunci. Salah satunya adalah String-frequency, metode ini telah mencapai hasil yang memuaskan dan telah digunakan secara luas. Ekstraksi kata kunci tradisional menggunakan algoritma TF-IDF, yang sederhana, cepat dan realistis. Ekstraksi kata kunci halaman web kebanyakan dilakukan untuk web kategori umum (general) **[3]**. Padahal kebutuhan untuk ekstraksi web yang spesifik untuk kategori anak juga tinggi, diantaranya untuk konten filtering, klasifikasi web anak, dan pencarian web anak. Penelitian ini telah melakukan eksperimen dan menghasilkan daftar sepuluh besar kata kunci dari web kategori kids, teens, dan mature teens [4].

2. Metode Penelitian

Ekstraksi kata kunci adalah pemilihan sekumpulan kecil kata atau frasa dari dokumen yang dapat menggambarkan makna dari dokumen tersebut. Metode penelitian dijabarkan pada gambar 1. Ekstraksi keyword dilakukan dengan mengekstrak seluruh teks yang berada dalam tag <meta> dengan atribut "class" bernilai "keyword" dari semua halaman web yang termasuk kedalam kategori ának dan umum **[5]**. Tahapan ekstraksi keyword adalah sebagai berikut:

a. Ekstrak keyword yang terdapat pada meta tag dari semua halaman web.

Pada tahap ini penulis melakukan proses *crawling* pada halaman web *http://dmoz-odp.org/* untuk mendapatkan menyalin semua informasi yang dibutuhkan sebagai data untuk proses klasifikasi halaman web. Penulis mengambil daftar halaman web yang ditujukan untuk anak dengan mengacu pada kategori *Kids & Teens Directory*. Data dengan label ['*Kids*'] digunakan sebagai data untuk kategori *kids* (anak

usia sampai dengan 12 tahun), label ['*Teens*'] untuk kategori *teens* (usia 13-15 tahun), dan label ['*Mature Teens*'] untuk kategori *mature teens* dengan usia 16-18 tahun.

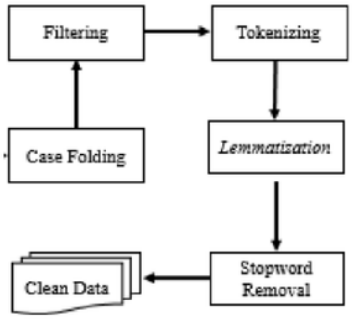
Sedangkan untuk halaman web yang ditujukan untuk umum penulis menggunakan kumpulan halaman web yang berada pada kategori *arts, business, computers, games, health, home, news, recreation, reference, regional, science, shopping, society, sport*. Jumlah data yang digunakan sebagai data set berjumlah adalah 311 url untuk kategori *kids*, 209 url untuk kategori *teens*, 339 url untuk kategori *mature teens* dan 300 url untuk kategori *general*. Data tersebut merupakan jumlah url yang berhasil di crawling dengan menggunakan aplikasi *webHarvy*. URL yang telah dikumpulkan tersebut akan diekstraksi dengan menggunakan crawler pada python.

Adapun proses crawling data web dilakukan melalui 3 tahapan: 1) pengumpulan link sub-direktori yang berisi URL dari setiap kategori; 2) Crawling URL berdasarkan kategori dengan menggunakan WebHarvy; dan 3) Ekstrak teks HTML dari URL dengan menggunakan beautifulSoap Phyton.

b. Hitung jumlah keyword yang berhasil dikumpulkan

Gambar 1 Metode penelitian

c. Lakukan *preprocessing* keyword pada tahap (b) sehingga mendapatkan keyword yang hanya terdiri dari satu kata. Terdapat 5 tahap preporcessing seperti yang dapat dilihat pada gambar 2. Adapun penjelasan untuk tiap tahapan adalah sebagai berikut :



Gambar 2 Tahapan Preproses

1. *Case Folding*

Case Folding yaitu mengubah semua huruf dalam teks menjadi huruf kecil.

2. *Filtering*

Filtering yaitu menghapus semua karakter kecuali huruf a-z.

3. *Tokenizing*

Tokenizing adalah sebuah proses untuk memilah isi teks sehingga menjadi satuan kata-kata.

4. *Lemmatization*

Lemmatization merupakan suatu proses untuk mengubah kata ke bentuk dasarnya.

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

5. *Stopword Removal*

Stopword removal adalah tahap untuk menghilangkan kata yang tidak penting seperti: saya, adalah, yang, dan sebagainya.

d. Hapus keyword yang duplikat, sehingga meninggalkan keyword yang bersifat *unique*.

Setelah keyword berhasil di ekstraksi, untuk mengetahui seberapa besar keyword mewakili isi dari keseluruhan dokumen maka akan dilakukan pengujian dengan menghitung jumlah term yang unik dan duplikat dari keyword terhadap 100 term terbanyak pada keseluruhan dokumen. Skema pengujian pada penelitian ini dapat dilihat pada gambar 3 berikut:

Gambar 3 Proses pengujian

3. Hasil dan Pembahasan

Dataset dan Skenario Pengujian Hasil

Kami mengumpulkan 1.159 halaman web yang diambil dari direktori D-MOZ menggunakan crawler Beautifulsoup dari Phyton. Dataset ini yang nantinya akan digunakan sebagai kumpulan data untuk ekstraksi kata kunci. Ada 4 kategori halaman web yang kami ambil untuk penelitian ini, yaitu *general, kids, teens, dan mature teens*. Untuk pengujian, kami mengambil secara random keyword pada masing-masing kategori secara manual untuk dijadikan sebagai groundtruth. Selanjutnya kami menghitung nilai precision recall hasil perbandingan antara ekstraksi keyword oleh sistem dengan groundtruth.

Pada penelitian ini, keyword akan diekstrak dari dokumen-dokumen dari kategori anak dan umum. Keyword didapatkan melalui proses crawling pada html dengan tag <meta> dengan class "keyword" yang pada bab sebelumnya sudah dijelaskan cara mendapatkannya.

Berikut adalah sepuluh keyword dari seluruh kategori dengan frekuensi paling banyak adalah : Game (106), Online (92), Kid (77), News (74), Free (49), Book (47), Scholarship (45), Science (43), dan Child (42)

Tabel 1 Top 100 keywords

Term	Word Frequency
Camp	84
Game	58
Online	50
Scholarship	50
News	47
Music	33

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

Term	Word Frequency
Science	31
Education	30
Art	28
Free	27

Tabel 2 berikut adalah hasil ekstraksi keyword untuk masing-masing kategori dengan menampilkan 10 keyword dengan frekuensi paling besar :

Tabel 2 Top 10 keywords

General		**Kids**		**Teens**		**Mature Teens**	
Keyword		**Keyword**		**Keyword**		**Keyword**	
Term	**Freq**	**Term**	**Freq**	**Term**	**Freq**	**Term**	**Freq**
Camp	83	Kid	77	Game	33	Scholarship	45
News	37	Game	53	Online	21	College	23
Summer	21	Craft	42	Learn	20	Student	16
Art	19	Child	38	School	17	Nuclear	15
Shop	17	Color	35	Free	15	Music	15
Online	17	Preschool	31	Teen	14	Aid	15
Theater	16	Book	30	Education	14	Science	14
Quote	15	Activity	28	Music	12	Picture	13

Dari hasil ekstraksi keyword yang ditampilkan pada tabel 2. Dapat diketahui bahwa dari 10 keyword teratas perbedaan keywords antar kategori signifikan karena hanya ada 3 term yang muncul pada lebih dari 1 dokumen. Term tersebut adalah online, game, dan music.

Untuk mengetahui seberapa besar keyword mewakili term dari keseluruhan dokumen maka akan dibandingkan dengan top 100 term dengan frekuensi paling banyak yang akan ditampilkan pada tabel 3 berikut:

Tabel 3 Top 10 Frequency of Term

General		**Kids**		**Teens**		**Mature Teens**	
Term in Doc		**Term in Doc**		**Term in Doc**		**Term in Doc**	
Term	**Freq**	**Term**	**Freq**	**Term**	**Freq**	**Term**	**Freq**
article	934	game	1000	site	289	time	690
news	803	kid	804	science	219	year	539
oct	670	math	669	page	210	student	447
world	611	learn	617	bird	205	world	434
post	585	world	590	quiz	204	school	366
art	568	people	562	time	176	pakistan	366
make	545	president	559	make	169	make	354
time	536	make	538	year	153	university	330

Dengan membandingkan 100 keyword dengan frequency paling banyak dengan 100 term dengan frequency paling banyak pada masing-masing kategori, maka didapatkan hasil jumlah keyword yang bersifat duplikat dan unik seperti pada tabel 4 berikut:

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

Tabel 4 Hasil Pengujian

	General	Kids	Teens	Mature Teens	Rata-Rata
Duplikat	32	29	35	26	30,5
Unik	68	71	65	74	69,5

Berdasarkan tabel 4 dapat diketahui bahwa hasil ekstraksi keyword dapat mewakili term dari seluruh dokumen dengan nilai rata-rata term yang duplikat sebesar 30,5% dengan nilai paling besar dihasilkan oleh kategori *Teens* yaitu sebanyak 35%.

4. Kesimpulan

Hasil ekstraksi keyword yang dilakukan pada penelitian ini dapat mengekstrak keyword dari kategori *General, kids, teens, dan mature teens* dengan keyword yang dihasilkan dari masing-masing kategori bersifat unik hanya terdapat 3 term yang muncul pada lebih dari 1 dokumen, sehingga setiap keyword dapat mewakili kategori masing-masing. Keyword juga dapat mewakili term dari keseluruhan dokumen dengan presentase rata-rata sebesar 30,5% untuk seluruh dokumen.

Referensi

[1] Y. Jin-sheng and M. Xin-wu, "Keyword Extraction from Chinese News Web Pages based on Multi-features," *Comput. Eng. Appl.*, vol. 2, pp. 1–9, 2013.

[2] Z. Kuo, X. Hui, and T. Jie, "Keyword Extraction Using Support Vector Machine," in *Proceedings of WAIM' 2006*, 2006, p. 85.

[3] L. Sujian, W. Houfeng, and Y. Shiwen, "News-Oriented Automatic Chinese Keyword Indexing," in *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, 2003, p. 92.

[4] J. Li, Q. Fan, and K. Zhang, "Keyword extraction based on tf/idf for Chinese news document," *Wuhan Univ. J. Nat. Sci.*, vol. 12, no. 5, pp. 917–921, 2007.

[5] A. Hulth, J. Karlgren, A. Jonsson, H. Bostrom, and L. Ask, "Automatic Keyword Extraction Using Domain Knowledge," in *Proceedings of Second International Conference on Computational Linguis- tics and Intelligent Text Processing*, 2001, p. 471.

comparison web kids

ORIGINALITY REPORT

6%	6%	4%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.scribd.com Internet Source	3%
2	id.123dok.com Internet Source	2%
3	henke.blogs.dsv.su.se Internet Source	2%

Exclude quotes	Off	Exclude matches	< 2%
Exclude bibliography	On		